MULTIMEDIA UNIVERSITY

STUDENT ID NO

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | |

# MULTIMEDIA UNIVERSITY

# FINAL EXAMINATION

**TRIMESTER 1, 2019/2020**

## TDS3301 – DATA MINING

(All sections / Groups)

15 October 2019
9:00 a.m. – 11:00 a.m.
(2 Hours)

---

**INSTRUCTIONS TO STUDENT**

1. This question paper consists of **4 pages** including cover page with **4 questions** only.
2. Answer **ALL** questions. All questions carry equal marks and the distributions of marks for each question is given.
3. Please write all your answers in the **Answer Booklet** provided.

## Question 1

a) There are many challenging issues in data mining work. Briefly describe **TWO** major issues in data mining. (2 marks)

b) *Outliers* are often discarded as noise. However, it can it be useful in numerous ways. State **TWO** applications where *outliers* can be useful. (2 marks)

c) The following table shows the midterm test results for a group of students where the rows show the marks they obtained for Question 1 (Qns.1) and Question 2 (Qns.2).

| Name | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Qns.1 | 3.5 | 4 | 5 | 6 | 4 | 5 | 7 | 3.5 | 9.5 | 3.5 |
| Qns.2 | 7 | 4 | 10 | 10 | 10 | 6 | 8 | 6 | 10 | 2.5 |

i. Determine the *mean*, *median,* and *mode* for both Qns.1 and Qns.2. (3 marks)
ii. Are there any *outliers* in the data? Provide proof for your answer. (2 marks)
iii. Are both data distribution skewed or normal? Based on this, what can be inferred regarding the questions in the midterm test? (1 mark)

**[TOTAL 10 MARKS]**

## Question 2

a) Data preprocessing can be categorized into four major tasks, namely *data cleaning*, *data integration*, *data reduction*, and *data transformation and data discretization*. Briefly describe the role of **EACH** task in supporting data mining. (4 marks)

b) The following table shows two attributes of weather data, outlook (sunny, rainy) and temperature (hot, cool).

| | Sunny | Rainy |
|---|---|---|
| Hot | 1505 | 330 |
| Cool | 225 | 940 |

i. Based on the given contingency table, calculate the $\chi^2$ of the data. (3 marks)
ii. At a significance level of 0.001, are the weather *outlook* and *temperature* correlated? (2 marks)
iii. Based on the correlation found in Question 2 (b) (ii), explain if the temperature can be a feature to predict the weather's outlook. (1 mark)

**Formulae:**

$\chi^2$ test, $\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$

Expected frequency, $E_i = \frac{count(A) \times count(B)}{n}$

**Continued ...**

Table of probabilities for Chi-squared distribution:

| df\significance | 0.1 | 0.05 | 0.01 | 0.005 | 0.001 |
|---|---|---|---|---|---|
| 1 | 2.706 | 3.841 | 6.635 | 7.879 | 10.828 |
| 2 | 4.605 | 5.991 | 9.210 | 10.597 | 13.816 |
| 3 | 6.251 | 7.815 | 11.345 | 12.838 | 16.266 |
| 4 | 7.779 | 9.488 | 13.277 | 14.860 | 18.467 |

**[TOTAL 10 MARKS]**

**Question 3**

a) State and briefly describe **ONE** method for classification. (1 mark)

b) List the step-by-step process of training a *Multi-layer Neural Network.* (2 marks)

c) The given table shows the records of vehicles from the Road Transport Department. Based on the available attributes, a *Naïve Bayesian Classifier* (NBC) and an *Artificial Neural Network* (ANN) were used to classify if the vehicles might be stolen or not given the attributes.

| Car plate | Colour | Type | Origin | Stolen? | | |
|---|---|---|---|---|---|---|
| | | | | Real | NBC | ANN |
| AX1234 | Red | Sports | Domestic | Y | Y | N |
| BPX1235 | Red | Sports | Domestic | N | Y | N |
| CCU1236 | Red | Sports | Domestic | Y | Y | Y |
| DT1237 | Yellow | Sports | Domestic | N | N | N |
| EAT1238 | Yellow | Sports | Imported | Y | Y | Y |
| FX1239 | Yellow | SUV | Imported | N | Y | N |
| HWA1240 | Yellow | SUV | Imported | Y | Y | Y |
| JJT1241 | Yellow | SUV | Domestic | N | N | Y |
| KLM1242 | Red | SUV | Imported | N | N | N |
| MAA1243 | Red | Sports | Imported | Y | Y | Y |

i. Create the confusion matrix for NBC and ANN (2 marks)

ii. Calculate the *accuracy, precision, recall,* and $F_1$*-score* for NBC and ANN. (4 marks)

iii. Which among the two classifiers should be chosen to predict whether a vehicle might be stolen? Please state the reason for the selection. (1 mark)

**Formulae:**

Precision, $P = \frac{TP}{TP+FP}$

Recall, $R = \frac{TP}{TP+FN}$

$F_1$-score, $F_1 = \frac{2PR}{P+R}$

**[TOTAL 10 MARKS]**

**Continued...**

**Question 4**

a) Explain the difference between *classification* and *clustering*. (1 mark)

b) The following table shows the *gender*, *height*, and *weight* of a sample population.

| ID | Gender | Height | Weight |
|---|---|---|---|
| 1 | M | 171 | 70 |
| 2 | F | 165 | 60 |
| 3 | F | 160 | 62 |
| 4 | F | 151 | 48 |
| 5 | M | 165 | 65 |
| 6 | M | 169 | 70 |

i. Normalize the numerical attributes to [0,1] using *min-max normalization*. (2 marks)

ii. Perform *k-means clustering* using the *Manhattan distance* ($L_1$-norm) as the distance measure to group the data into 2 clusters. (Execute **TWO** iterations using Object 1 and Object 2 as initial cluster centers and state which cluster does each of the objects belong to). (6 marks)

iii. Based on the clustering outcome in (ii), can gender be predicted given height and weight? Give reasons for your answer. (1 mark)

**[TOTAL 10 MARKS]**

Min-max normalization, $v' = \frac{v - min_A}{max_A - min_A}(new_max_A - new_min_A) + new_min_A$

Minkowski distance, $d(i,j) = \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p}$

**End of page.**